Big Data for Prediction: Patent Analysis

Mirjana Pejić Bach

*Faculty of Economics and Business, University of Zagreb, Croatia*

Jasmina Pivar

*Faculty of Economics and Business, University of Zagreb, Croatia*

Živko Krstić

*Crossing technologies, Croatia*

**ABSTRACT**

*Usage of big data technologies for prediction in various fields, such as retailing, marketing and social media, lures attention of different stakeholders. The reasons are related to the potentials of big data, which allows learning from past behavior, discovering patterns and values, and optimizing business processes based on new insights from large databases. However, in order to utilize the potentials of big data fully, its stakeholders need to gather insight in new trends in this area. Patent analysis is an efficient methodology used for technological insight for numerous technologies. The goal of this paper is to use patent analysis is order to gain technological insight in the area of big data technologies usage for prediction. This is done by: (i) exploring the timeline and geographic distribution of patents of big data solutions for prediction, (ii) exploring the most active assignees of patents of big data solutions for prediction,, (iii) detecting the type of technolgies protected by patents of big data solutions for prediction, using the International Patent Classification system, and (iv) performing text-mining analysis to discover the topics emerging most often in abstracts of patents of big data solutions for prediction.*

Keywords: Big Data, Prediction, Technological Field, Patent, Simple Patent, Family, International Patent Classification, PatSeer, Patent Analysis, Association Rules, Text Mining

**INTRODUCTIOn**

Patent databases are an abundant and important source of information about the particular technical field, and patent analysis has been proven as effective tool for decision makers who seek for a comprehensive overview of different technologies’ topics, such as big data technologies (Madani and Weber, 2016). Decision makers may want to understand relevant trends, to spot new technologies in particular area or to estimate the importance of the emerging new technologies. Moreover, patent information is a relevant source for those who want to get familiar with key players of a particular technology, or to learn about their productivity and patenting behavior.

Big data technologies have attracted lots of attention due to their ability to analyze large amounts of various data sources, and extract useful information from them. Recently, big data technologies have become not only a methodology for analyzing the current situation, but are also used as tools for prediction in various fields, such as retailing, marketing and social media (e.g. Bradlow et al., 2017; Miah, Vu, Gammack and McGrath, 2017; Shirdastian et al., 2017).

Goal of this chapter is to analyze and help to understand patents related to big data for prediction. The paper will provide answers to the following questions that are of interest to big data inventors and investors: (1) What is the timeline of patents of big data solutions for prediction?; (2) Who are assignees of patents of big data solutions for prediction, and what is their geographic origin?; (3) What are the most frequent IPC patent areas of patents of big data solutions for prediction?, (4) What are the most often topics of patents of big data solutions for prediction? Answers to these questions will provide useful guidance related to competitiveness and new trends that emerge in the usage of big data technologies for prediction. Additional goal of this paper is to assess the usability of several data mining and text mining methods for the purpose of patent analysis, specifically association analysis of IPC patent areas, key-terms extraction and clustering. For this purpose, Statistica Text Miner 13.0, and Provalis Wordstat 8.0 has been used.

The chapter consists of the following sections. After the introduction, the second section presents the background of the research, encompassing the notion of big data, usage of big data for prediction, and usage of patent analysis. The third section describes the methodology used. The results of the analysis are presented in the fourth section. Finally, the last section is used to synthesise findings, present limitations, and future research directions of the chapter.

**BACKGROUND**

**Big Data and Predictive Analytics**

Big data has become an exciting field of study for practitioners and researchers, due to the need to adapt to the emergence of huge databases (Parr Rud, 2011). Each of them has different focus and concerns in this area, which yielded various definitions and descriptions of big data. Practitioners, such as consulting companies and multinational corporations, define big data by mainly focusing on the technology necessary to handle such data. For example, the National Institute of Standards and Technology describes it as data that exceed capacity or capability of conventional systems and “require a scalable architecture for efficient storage, manipulation and analysis” (NIST, 2017, p. 8). On the other hand, scientists describe big data as the phenomenon related to various characteristics of data generated by different actions, e.g. social media and business transactions. Boyd and Crawford (2012, p. 662) define big data as “cultural, technological, and scholarly phenomenon that rests on the interplay of technology, analysis and mythology”. Furthermore, scientists often use following three characteristics in order to describe big data: Volume, Variety and Velocity. Volume describes the large amount of data that depends on the type of data, time and industry, which “make it impractical to define a specific threshold for big data volumes” (Gandomi and Haider, 2014, p. 137). Variety refers to the various types of data including structured, semi-structured and unstructured data (Chen et al., 2014), while velocity relates to the rapid and timely conducted data collection and data analysis (Chen et al., 2014; Dmitriyev et al., 2015; Vera-Baquero et al., 2015).

Harnessing big data is believed to result in more efficient and effective operations (Günther et al., 2017). Moreover, big data is being perceived as support for decision-making (Sharma and Kankanhalli, 2014) or as a source of business opportunities (McAfee and Brynjolfsson, 2012; Gandomi and Haider, 2014). Günther et al. (2017) stress out that continuous interaction between work practices, organizational models and stakeholder interests are the prerequisites for the successful usage of big data. Big data analytics is the main source of value generated by big data technologies, since it allows generation of the new knowledge from huge databases, which only recently emerged as a possibility. Big data analytics refers to exploitation of algorithms that can process a large volume of various types of data at increasing speeds, which can be classified into following groups: text analytics or text mining, multimedia analytics, social media analytics and predictive analytics:

* Text analytics denotes techniques for extraction of useful information and knowledge from unstructured textual data (e.g. business documents, emails, social media). Text mining is primarily based on natural language processing (NLP) which enables computational text analysis, interpretation and generation (Chen, Chiang and Storey, 2012; Chen et al., 2014; Gandomi and Haider, 2015). Examples of common NLP-based techniques used in text analytics are text summarization techniques, opinion mining, clustering, and so on.
* Multimedia data analytics refers to information extraction from unstructured audio, images and video streams data. The transcript-based approach and phonetic-based approach are two common technological approaches to audio analytics (Gandomi and Haider, 2015). Video analytics or video content analysis refers to various techniques for analyzing and extracting information from video data.
* Social media analytics encompasses techniques for analyzing both structured and unstructured data generated by social media (Chen et al., 2012; Gandomi and Haider, 2015). Social media analytics is classified into content-based analytics and link-based analytics. Content-based analytics refers to usage of text; video and audio analytics for analyzing data generated by users of social media, such as images, reviews and so on. Link-based analytics is focused at structure of social networks and relationships among entities that participate in networks. For example, community detection techniques can be used to uncover behavioral patterns and predict properties of certain network. Additionally, participants’ influence or strength of connections in networks can be evaluated by using so-called social influence analysis. Similarly, link prediction strives to predict future linkages between entities in a network.
* Predictive analytics use both quantitative and qualitative approaches to learn from past behavior, uncover patterns in data and to optimize business processes based on new insights. It usually refers to the application of statistical techniques, data mining and machine learning algorithms to extract information and knowledge from structured data. Common goals of various predictive analytics approaches are to found patterns in data and to explore relationships in data.

Business application domains that current focus for big data and predictive analytics are retail, marketing, and social media. Bradlow et al. (2017) examine the opportunities of using data about customers, products, time, location and channel for the purpose of decision making in retailing, using Bayesian techniques on large dataset. Miah et al. (2017) propose the method for analysing unstructured data, geo-tagged photos uploaded by tourists to social media, to support strategic decision-making in tourism destination management. Salehan and Kim (2016, p. 31) suggest an approach for development of “scalable, automated systems for sorting and classification of big online consumer reviews data, which will benefit both vendors and consumers”. Yi and Wang (2017, p. 188) presented “a big data analytics based fault prediction approach for shop floor scheduling”. Latent semantic analysis and the support vector machine were used to examine the sentiments toward a brand to identify the reasons for positive or negative sentiments on social media (Shirdastian et al., 2017).

Some authors discussed application areas that predictive analytics using big data will greatly influence in future. Akter and Wamba (2016) review usage of big data analytics in e-commerce. They concluded that main application areas of big data analytics in e-commerce are personalization, dynamic pricing, customer service, supply chain visibility, security and fraud detection, as well as predicting individual customer’s theoretical values to company, to predict sales patterns, to forecast and determine inventory requirements and to predict consumer preferences and behavior. Big data analytics attracted attention in various areas such as logistics and supply chain management (Waller and Fawcett, 2013), cyber-physical systems (Lee et al., 2015), auditing (Geep et al., 2018), cognitive computing (Garret, 2014) helath care services (Wu et al., 2016), cybersecurity (Rassam et al., 2017).

**Patent Analysis for Decision Making**

Decision makers who seek for a comprehensive overview of different technology topics in a technical field of interest may rely on patent analyzes, which often utilizes text mining (Pejić Bach et al., 2017). Madani and Weber (2016) analyze the evolution of patent analysis, focusing to text mining. Brügmann et al. (2014) present workbench for intelligent patent document analysis, which includes modules for summarization, entity recognition, segmentation, lexical chain identification and claim-description alignment. For example, Kim et al. (2016) use the semantic patent topic analysis-based bibliometric method to generate patent development maps related to 3D printing technologies. Altunas et al. (2014) analyzed patent documents by using weighted association rules that recognise the different importance of protected technical content based on following criterion: commercial significance and technological impact. Patent lanes developed regard semantic similarities, which can be seen as the deployment of patent clusters, were suggested by Niemann et al. (2017) in order to describe the development of a technological field in the course of time. Han et al. (2017) presented usage of natural language processing technologies to extract concepts and patent similarity assessments, and to support content-oriented visualisation.

Valuable insights lie in patent citations which analysis can reveal patterns of knowledge spillover and diffusion of information between different stakeholders such as countries, universities and companies. Patent citation analysis reveals its applicability across different technical fields that serve the creation of technology (Sharma and Tripathi, 2017). Kyebambe et al. (2017) used supervised learning methods to forecast emerging technologies. Furhermore, Kim and Bae (2017) suggested a three-step methodology for technology forecasting. The first step is to cluster patent documents based on cooperative patent classification. The second step is to examine the combination of cooperative patent classification of each derived clusters. The final step is to determine which clusters are promising based on analysis of patent indicators such as citations, triadic patent families as well as independent patent claims. Song et al. (2018) used a bibliographic coupling to patents to produce a list of outlier patents, developed the technological and market measures to evaluate them and determined promising technologies based on the developed measures.

Patents can be searched and analyzed by using numerous patent databases or platforms. Patent databases can be divided into national databases and world databases. Examples of national databases are United States Patent and Trademark Office (USPTO) patent database, Canadian Intellectual Property Office patent database, Australian patent database - AutPat or DEPATISnet, which contains patents from the German Patent and Trade Mark Office. Patent databases that contain patent documents from around the world are Espacenet, Google Patents, The Lens, Patentscope, which provides access to international Patent Cooperation Treaty (PCT) applications, and OECD Patent Database that contains data on patent applications to the European Patent Office - EPO and USPTO. Commercial patent platforms allow advanced patent search and analysis such as patent network analysis or citation analysis. Examples of commercial patent platforms are PatSeer, Clearstone Elements, PatentCloud, LifeQuest, Derwent Innovation by ClarivateAnalytics, Total Patent One by Lexis Nexis and Octimine.

**METHODOLOGY**

Patents from the PatSeer database related to big data usage for prediction analytics from 2013 to 13 October 2017 are analyzed, using the longitudinal approach in combination with text mining techniques. The patent analysis consists four phases related to (i) the patent search and selection, (ii) timeline, geographic origin and patents assignees analysis, (iii) patents analysis according to IPC system patent area, and (iv) text mining.

**Phase One: Patent Search and Selection**

A patent, in general, is an exclusive right granted for an invention to exclude others from making, using, or vending the patented invention without the patent owner's permission. Each patent’s information or so-called meta-data of patents are provided in the form of highly structured documents. Patent documents usually contain following patent’s data: title, abstract or description, publication or issue year, filing/application year, priority country, assignee country, The International Patent Classification codes, The Cooperative Patent Classification CPC codes, File Index codes, backward/forward citations and so on. Analysis of patents’ documents containing all of these data sheds light on a technical area of interest and can serve to stakeholders in their decision-making. Patent databases should provide accurate data in comprehensible format and deliver data promptly (Madani and Weber, 2016) in order to be relevant and valuable for decision makers.

PatSeer is an online patent database storing the patents in the forms of simple patent families. PatSeer is available in several editions: Lite, Standard, Premier, Pro, Explorer and Projects Edition. Authors used Lite Edition to conduct a preliminary search of the simple patent families to detect the patents related to big data for prediction. In general, Lite Edition is used to search the worldwide patent database and allows users to manage and save search strings, to narrow down search results by using filters, as well as to extract data in excel format. Therefore, authors used PatSeer solution for searching and extracting patent data only. Other PatSeer’s Editions offer more capabilities in comparison to Lite Edition. For example, PatSeer Pro allows advanced patent analysis such as patent network analysis with semantic spatial-mapping, to conduct citation analysis, text clustering and more.

The PatSeer database was searched on 13 October 2017 by using the search string search string (TA: (data AND (predict OR prediction OR forecasting OR forecast OR prognosis OR prognosticate OR foresight OR foresee))), with an option for searching simple patent families. Authors found 316 of records for simple family families in total. Among these records, 296 simple patent families had status “active” at the time of the search. Therefore, a patent analysis of the 296 simple patent families related to big data for prediction was conducted to achieve the goal of this research.

**Phase Two: Timeline, Geographic Origin and Patents Assignees Analysis**

Authors performed an extensive analysis of timeline, geographic origin and current assignees in order to detect which of them were most active in patenting technical content related to big data for prediction. A current assignee is an entity, organization or individual, inventor, that has the property right to the patent (Sinha and Pandurangi, 2015).

**Phase Three: Patents according to IPC system patent area**

Authors analyzed the protected technical content of big data for prediction simple patent families, using International Patent Classification (IPC) system established in 1971 by the Strasbourg Agreement, used in more than 100 countries worldwide. The IPC describes technical knowledge by using the systematic and hierarchical classification, which includes section, class, subclass, group and subgroup (WIPO, 2017). In this research, the analysis of the active simple patent families related to big data usage for prediction according to the sections, subclasses and groups will be conducted. In order to determine whether the technical content of the selected simple patent families is heterogeneous or homogeneous, authors use association rules.

**Phase Four: Text Mining Patent Analysis**

Text mining approach was utilised in order to detect the topics emerging most often in abstracts of simple patent families related to big data solutions for predictive analytics. Software WordStat Provalis was used for text mining. First, phrases of maximum five words, which occur in more than five simple patent abstracts, are extracted. Second, extracted phrases were used to conduct cluster analysis in order to detect which topics occur together. Cluster analysis of phrases was conducted using of average-linkage hierarchical clustering algorithm, which creates clusters from a similarity matrix (Everitt et al., 2011). The distance between two clusters is the average distance between each observation in one cluster to every observation in the other. This method is also called Unweighted Pair Group Mean Averaging. For example, distance between clusters “A” and “B’’ refers to average length of each arrow connecting observations within the clusters (Figure 1) as expressed in Formula 1.

*Figure 1. Average linkage method*

B

A

*Source: (Authors)*

*Formula 1. Distance between clusters – Average linkage method*

|  |  |
| --- | --- |
|  | (1) |

Notation:

A1, A2,..., Ak = Observations from cluster A

B1, B2,..., Bl = Observations from cluster B

d (a,b) = Distance between a cluster with observation vector *a* and a cluster with observation vector *b*

The cluster analysis was conducted by using Jaccard's coefficient as a similarity measure. Jaccard’s coefficient determined the association between two phrases that occur together in simple patent abstract. The result is represented by the dendogram. Single-word clusters were hidden from the dendrogram to simplify the use of the dendrogram and being able to focus only on the strongest associations of meaningful phrases. Since a dendrogram determines only the temporal order of the branching sequence, the sequence of phrases cannot be seen as a linear representation of those distances. That means that any cluster can be rotated around branches on the dendrogram without affecting its meaning. For that reason, authors used proximity plots generated in WordStat Provalis software in order to represent the distance between most frequent phrases to all other phrases. In proximity plot, phrases that often tend to appear near selected phrase are shown on the top of the plot. In addition, network graphs were used in order to represent the relationships between phrases by lines connecting those phrasest.

**RESULTS**

In this part of the chapter, patent analysis results are presented as following: timeline, geographic origin and patents assignees of related to big data for prediction, the result of patents analysis according to IPC system patent area and results of the text mining patent analysis.

**Timeline, Geographic and Assignee Patent Analysis**

In order to provide answers to when, where and who pursues protection of big data analytics solutions for predictive analysis, the timeline, geographic and assignee analysis was conducted.

Table 1 represents the timeline for the period between 2013 and October 2017, and geographic origin of simple patent families.

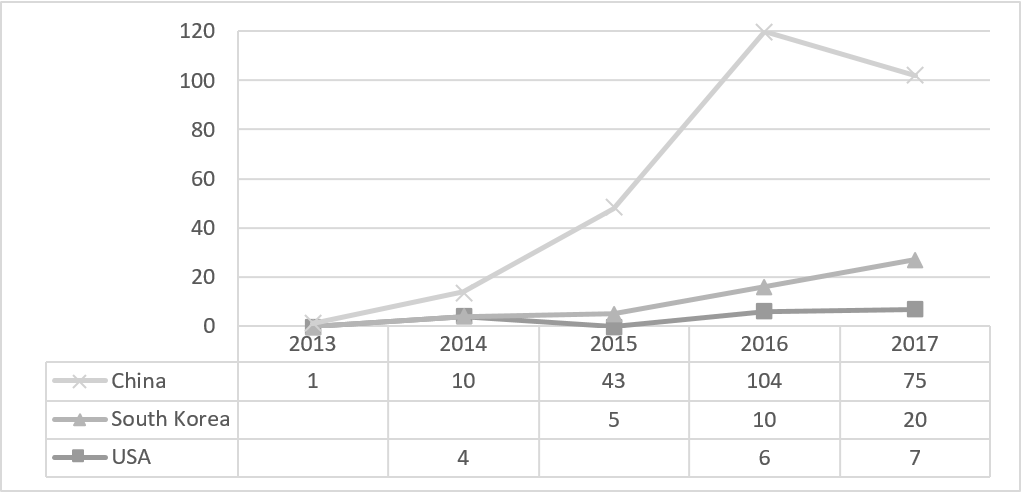
*Table 1. Number of big data for prediction simple patent families per publication/issue year and priority country (from 2013 to 13th* *October 2017)*

|  |  |  |
| --- | --- | --- |
| **Publication / Issue Year** | **No. of Simple Patent Families** | **% of Total No. of Simple Patent Families** |
| **Timeline** | | |
| **2013** | 1 | 0% |
| **2014** | 16 | 5 % |
| **2015** | 50 | 17% |
| **2016** | 122 | 41% |
| **October 2017** | 107 | 36% |
| **Total** | 296 | 100.00% |
| **Country of Origin** | | |
| **Priority Country** | **No. of Simple Patent Families** | **% of Total No. of Simple Patent Families** |
| **China (CH)** | 233 | 79% |
| **South Korea (KR)** | 35 | 12% |
| **United States of America (USA)** | 17 | 6% |
| **India (IN)** | 4 | 1 % |
| **Taiwan (TW)** | 3 | 1% |
| **Japan (JP)** | 1 | 0% |
| **None** | 3 | 1% |
| **Total** | 296 | 100.00% |

*Source: (Authors, PatSeer, 13th October 2017)*

Most of the most of the assignees related to big data for prediction are spread across China and South Korea. Figure 2 provides details on the timeline and geographic origin of simple patent families according to priority countries for the period between 2013 and October 2017.

*Figure 2. Number of big data for prediction simple patent families per priority country (from 2013 to 13th October 2017)*



*Source: (Authors, PatSeer, 13th October 2017)*

Table 2 provides details on the number of simple patent families related to big data for prediction according to current assignees and countries, which indicates that all organizations with more than 5 patents come from China.

*Table 2. Number of big data for prediction simple patent families according to current assignee and country (from 2013 to 13th October 2017)*

|  |  |  |  |
| --- | --- | --- | --- |
| **Current Assignee** | **Country** | **No. of Simple Patent Families** | **% of Total No. of Simple Patent Families** |
| State Grid Corporation | China | 22 | 7.4% |
| Inspur Group | China | 7 | 2.4% |
| Nanjing University | China | 7 | 2.4% |
| Business Big Data | China | 5 | 1.7% |
| Hohai University | China | 5 | 1.7% |
| Other | - | 250 | 84.5% |
| **Total** | | 296 | 100.00% |

*Source: (Authors, PatSeer, 13th October 2017)*

**Patents According to IPC System Patent Area**

Majority of the simple patent families were assigned to more than one IPC’s main groups or sub-groups. A simple patent family is usually registered under multiple ICR codes, so the total number of ICR codes (561 codes) is larger than the number of simple patent families examined (296 simple patent families), which indicates that one simple patent family is registered to approximately two IPC’s main groups or sub-groups on average. Observed simple patent families were registered under following seven five IPC sections: A Human Necessities; B Performing Operations; Transporting; C Chemistry, Metallurgy; E Fixed Constructions, F Mechanical Engineering; Lighting; Heating; Weapons; Blasting Engines or Pumps, G Physics and H Electricity.

Table 3 presents the number of big data for prediction simple patent families according to the IPC system – Sub-class level, that occur in more than 10 simple patent families. Among classes assigned to 296 simple patent families are computing, calculating or counting instruments such as G06Q - Analogue computers (228 times), G06F - Electrical digital data processing (132 times) and G06N - Computer systems based on specific computational models. Additionally, simple patent families that were registered as an electric communication technique were mostly related to the sub-class H04L - Transmission of digital information.

*Table 3. Number of big data for prediction simple patent families according to the IPC system – Sub-class level (>10 simple patent families)*

|  |  |  |
| --- | --- | --- |
| **Subclass** | **Description** | **No. of Simple Patent Families** |
| **A Human Necessities** | | |
| **A61B** | Medical diagnosis, surgery and identification | 12 |
| **G Physics** | | |
| **G06Q** | Analogue computers | 228 |
| **G06F** | Electrical digital data processing | 132 |
| **G06N** | Computer systems based on specific computational models | 26 |
| **G08G** | Traffic control systems | 16 |
| **G06K** | Instruments for recognition and presentation of data | 14 |
| **G08B** | Signaling or calling systems - order telegraphs, alarm systems | 13 |
| **G05B** | Monitoring or testing arrangements/elements for control systems | 12 |
| **H Electricity** | | |
| **H04L** | Transmission of digital information | 27 |
| **Other** |  | 76 |
| **Total** |  | 561 |

*Source: (Authors,* *PatSeer, 13th October 2017)*

Table 4 presents simple patent families according to IPC main group and sub-group level. Data processing systems or methods adapted forecasting or optimization was the most frequent IPC’s group. A substantial number were related to administrative, financial, managerial or supervisory purposes – IPC’s group G06F17/30 (62 simple patent families). Additionally, 40 of 228 simple patent families that were registered for electricity, gas or water supply purposes.

*Table 4. Number of simple patent families related to big data for prediction according to the IPC system – Main group/Sub-group level (>10 simple patent families)*

|  |  |  |
| --- | --- | --- |
| **Main/Sub Group** | **Description** | **No. of Simple Patent Families** |
| **G06 Physics - Computing, calculating and counting instruments** | | |
| **G06F Digital computing or data processing equipment or methods for:** | | |
| **G06F17/30** | Administrativec, financial, managerial, supervisory purposes | 62 |
| **G06F19/00** | Specific applications | 28 |
| **G06Q Data processing systems or methods specially adapter for:** | | |
| **G06Q10/04** | Forecasting or optimization | 64 |
| **G06Q50/06** | Electricity, gas or water supply | 40 |
| **G06Q10/06** | Resources, enterprise planning, organizational model | 22 |
| **G06Q30/02** | Marketing, e.g. Buyer profiling, price estimation | 19 |
| **G06Q50/26** | Government or public services | 12 |
| **G06Q50/10** | Services | 11 |
| **H04 – Electricity - Electric communication technique** | | |
| **H04L Transmission of digital information** | | |
| **H04L29/08** | Control procedure, e.g. Data link level control procedure | 12 |
| **Other** | | 291 |
| **Total** | | 561 |

*Source: (Authors, PatSeer, 13th October 2017)*

**Co-occurrence of IPC areas**

In order to detect relationships between IPC codes, association rule analysis was conducted. IPC’s main group or sub-group code is considered as an item, and each record of a simple patent family is considered as a transaction. Due to the heterogeneity of IPC codes, task for finding association rules was non-trivial and association rules between different IPCs’ main groups or sub-groups level codes were challenging to detect. Therefore, minimal support and confidence at 1% level was set, which resulted in 39 association rules. Table 5 shows only rules with the minimal support of 2% and minimal correlation of 10%, which reveals that the simple patent families registered as data processing systems or methods for forecasting or optimization were specially adapted for electricity, gas or water supply purposes in 10.47% of the total number of simple patent families (Rule G06Q10/04 🡪 G06Q50/06). Data processing systems or methods for resources management were specially adapted for electricity, gas or water supply purposes in 2.70% of the total number of simple patent families (Rule G06Q10/06 🡪 G06Q50/06).

*Table 5. Summary of association rules - Min. support = >2%, Min. confidence = >2%, Min. correlation = 10%*

|  |  |  |  |
| --- | --- | --- | --- |
| **Body – Description (application area or method)** | **Head – Description (application area or method)** | **Support/ Confidence** | |
| G06Q10/04 - forecasting method | G06Q50/06 - energy supply | 10% | 48% |
| G06Q50/06 - energy supply | G06Q10/04 - forecasting method | 10% | 78% |
| G06Q10/06 - enterprise resources planning | G06Q50/06 - energy supply | 3% | 36% |
| G06Q50/06 - energy supply | G06Q10/06 - enterprise resources planning | 3% | 20% |
| G06F17/30 - finance/management | G06Q10/04 - forecasting method | 2% | 11% |
| G06Q10/04 - forecasting method | G06F17/30 - finance/management | 2% | 11% |
| G06Q10/04 - forecasting method | G06Q50/26 - government/public services | 2% | 9% |
| G06Q10/04 - forecasting method | G06Q50/26 - government/public services | 2% | 9% |
| G06Q10/06 - enterprise resources planning | G06Q10/04 - forecasting method | 2% | 27% |
| G06Q50/26 - government/public services area | G06Q10/04 - forecasting method | 2% | 50% |

*Source: (Authors, PatSeer, 13th October 2017; Statistica Text Miner)*

**Patent topics**

In order to detect most frequent topics of the simple patent families’ abstracts, authors used the phrase extraction process combined with the cluster analysis conducted by Wordstat Provalis software. Authors detected following most frequent phrases: real-time, data mining, early warning and neural networks.

Table 6 shows most frequent phrases in patent applications with the frequency of occurrence ≥ 5. Column TF\*IDF of Table 7 contains values of metrics for a phrase’s importance. The Term Frequency-Inverse Document Frequency (TF-IDF ) is a metric that helps to estimate how important is a phrase in a whole collection of documents (e.g. abstracts of all analyzed patents in a certain area) and not only in a particular document (e.g. abstract of only one patent). Therefore, for this chapter, TF-IDF is a metric that helps authors to estimate how important is a phrase in a whole collection of analyzed patents. Specifically, for this research, the collection of patents refers to patents’ abstracts.

Reason for using TF-IDF metric is that common words usually appear several times in a document (an abstract of certain patent), but they are not important as key-phrases to be searched or indexed. Term Frequency measures how frequently a phrase occurs in an abstract of patent. The Term Frequency value for the certain phrase "p" in the certain patent’s abstract is defined as the ratio between the frequency of phrase "p" in the patent’s abstract, and the total number of phrases in the same patent’s abstract. Furthermore, Inverse Document Frequency measures how important is a certain phrase "p" concerning the whole collection of patents’ abstracts. The IDF for a given keyword "p" in the collection of patents is calculated as the logarithm of the ratio between the total number of patents’ abstracts in a collection and is the number of abstracts in which the phrase "p" appears. Finally, the product of TF and IDF value gives its TF-IDF value for a certain phrase "p". Therefore, a phrase that has higher TF-IDF values is of higher importance. Phrases that are most important in the whole collection of patents related to big data for prediction, indicated by their TF-IDF values, are real-time (TF-IDF value 79.7), data mining (48.9), early warning (58.5) and neural networks (63.2).

*Table 6. Most frequent phrases in patent applications (>5% of Cases)*

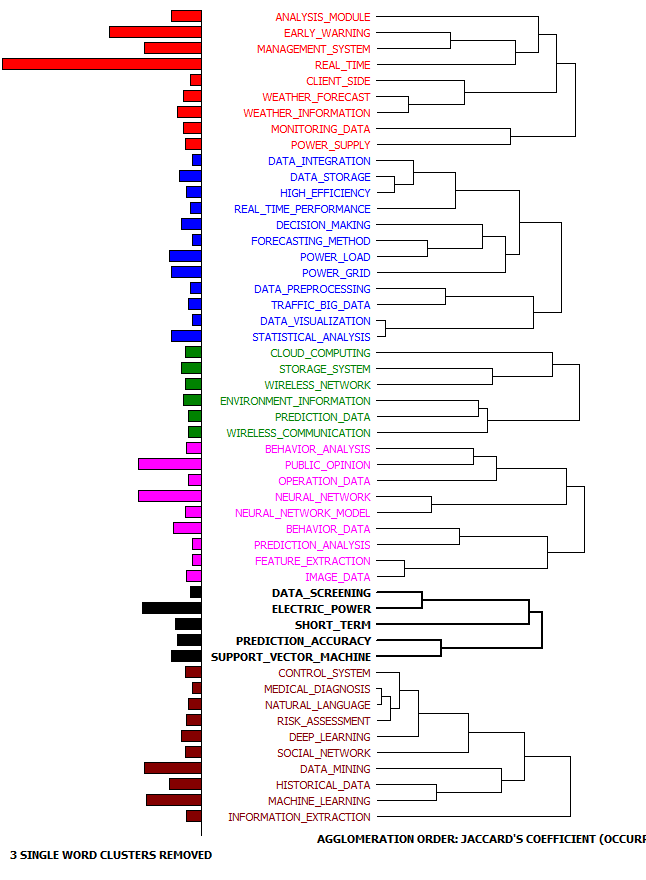
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Phrase** | **Frequency** | **No. of Cases** | **% Cases** | **TF - IDF** |
| real time | 109 | 55 | 18.58% | 79.7 |
| data mining | 52 | 34 | 11.49% | 48.9 |
| early warning | 58 | 29 | 9.80% | 58.5 |
| neural network | 57 | 23 | 7.77% | 63.2 |
| management system | 30 | 18 | 6.08% | 36.5 |
| machine learning | 29 | 16 | 5.41% | 36.7 |
| historical data | 17 | 15 | 5.07% | 22.0 |
| data platform | 25 | 15 | 5.07% | 32.4 |

*Source: (Authors by using* *WordStat Provalis software)*

Figure 3 presents the results of the cluster analysis that identified six groups of topics regard simple patent families related to big data for prediction.

* Cluster 1 includes 28 simple families patents’ abstracts, with the co-occuring phrases: real-time systems used for weather forecasting to provide weather information to a client-side; early warning management system based on monitoring data for managing power supply.
* Cluster 2 includes 10 simple families patents’ abstracts with the co-occuring phrases: data analysis supported by efficient database technologies such as managing power grid based on power load forecasting method or preprocessing of big traffic data.
* Cluster 3 includes 6 simple families patents’ abstracts with the co-occuring phrases: environment information and prediction data supported by wireless communication; storage systems and wireless communication supported by cloud computing and wireless networks.
* Cluster 4 includes 11 simple families patents’ abstracts with the co-occuring phrases: predicting and monitoring public opinion, and analyzing behavior data by using feature extraction and neural networks.
* Cluster 5 includes from 12 simple families patents’ abstracts with the co-occuring phrases: using support vector machine to increase prediction accuracy.
* Cluster 6 includes 12 simple families patents’ abstracts with the co-occuring phrases: information extraction based on data mining and machine learning to analyze historical data; information extraction based on deep learning for control systems and risk assessment, as well as a medical diagnosis based on natural language processing.

*Figure 3. Cluster dendrogram of phrases that occur in most frequent phrases*



*Source: (Authors by using WordStat Provalis software)*

In a dendrogram, the phrases (keywords) that co-occur tend to appear near each other but dendrogram determines only the temporal order of the branching sequence. For that reason, reading dendrograms is not intuitive or very easy. Therefore, authors used proximity plots generated to detect phrases that often tend to appear near selected phrase (Figure 4). Such phrases are shown on the top of the plot.

Figure 4 presents four proximity plots indicating which phrases occur the most often with the most frequent and most important phrases: real-time, data-mining, early warning and neural network. Authors found following:

* The phrases that occur the most often with the phrase *real-time* are mostly related to *data analysis* such as historical data, management systems, real-time performance and monitoring data; methods and techniques for data analysis such as statistical analysis, neural networks, machine learning or data visualization, as well as specific purposes such as traffic big data, power supply, risk assessment, social networks or behavior analysis.
* The phrases that occur the most often with the phrase *data mining* are mostly related to the phrase historical data, methods and techniques of data analysis such as machine learning, natural language or deep learning, and applications such as medical diagnosis, risk assessment or control systems.
* The phrases that occur the most often with the phrase *early-warning* indicate general technical parts of early warning systems such as management system, an analysis module, real-time and client side, as well as particular purposes of early-warning systems such as weather forecasting and power supply management. The phrase is also related to phrases indicating source or type of data used or generated by early-warning systems such as monitoring data, weather information, environment information.
* The phrase *neural network* occurs the most often with the phrase neural network model. Other phrases that occur with the phrase neural network indicate its’s specific application areas such as power load, behavior analysis, weather forecast, feature extraction or medical diagnosis. Additionally, types of data analyzed by neural networks are indicated by phrases historical data and behavior data.

Furthermore, the connections between keywords – phrases are visualized by using a network graph that allows us to explore relationships, to detect underlying patterns and structures of co-occurrences. Network graph was generated for each of the six clusters in the dendrogram. Elements are represented as a node while their relationships are represented as lines connecting those nodes. Figure 5 presents six network graphs indicating which phrases co-occurred most often within each of the cluster.

*Figure 4. Proximity plots of phrases that occur in more than 20 patent applications*

|  |  |
| --- | --- |
|  |  |
|  |  |

*Source: (Authors by using* *WordStat Provalis software*

*Figure 5. Network graphs of phrases that occur most frequent*

|  |  |
| --- | --- |
| Cluster 1 | Cluster 2 |
| Cluster 3 | Cluster 4 |
| Cluster 5 | Cluster 6 |

*Source: (Authors by using WordStat Provalis software)*

**FUTURE RESEARCH DIRECTIONS**

This chapter provides an outlook to the possible questions that can be answered for the investors and inventors interested in big data solutions for predictive analytics. Patent analysis can provide answer to the most basic questions, relating to when and where most of the patenting was conducted, by whom and in which areas. Therefore, future research directions are provided as the answers to these questions.

*When?*

Analysis indicate that area of big data usage for predictive analytics emerged recently. Only one simple patent family related to big data for prediction was registered in 2013. After that period, the number of simple patent families increases rapidly, with 122 simple patent families registered in 2016 and 107 simple patent families registered in 2017, until October. The emerging trend is expected to continue in the period of at least several years.

*Where?*

China is the leading country in patenting activities related to big data for prediction. Chinese organizations began publishing patents related to this technical area in 2013. South Korea began publishing big data for prediction patents two years later, in 2015. Among other countries, only India, Japan and Taiwan published big data for prediction patents.

*Who?*

The organization that registered the most substantial number of simple patent families related to big data for prediction in the observed period is State Grid Corporation registered in China (227 simple patent families ). Inspur Group (7 simple patent families) and Nanjing University (7 simple patent families) were active assignees from China as well. Kim Seung Chan, the inventor, registered three simple patent families, which makes him being the only individual on a list of first ten assignees of the area of interest. Other organizations that registered a more substantial number of simple patent families are companies such as Business Big Data, NAT Computer Network Information Security, Shanghai Fuli Information Technology and academic institutions such as Hohai University, Beijing Jiaotong University and the University of South China.

Patenting applications related to big data and prediction have been followed trends that are present in patent activities worldwide generally. According to patenting indicators for 2016, published by World Intellectual Property Organization (2017), China is the largest contributor in number of filing. The State Intellectual Property Office of The People’s Republic of China (SIPO) received more than 1.3 million patent applications in 2016, which was more than the European Patent Office, the United States Patent and Trademark Office, the Japan Patent Office and the Korean Intellectual Property Office received together. Many of patents are related to new technological content in computing, medical technology, semiconductors, and so on. Reasons, why patenting activities in China have been growing, are following. In 2012, China’s government set the goal regard growth of all type of patenting activities. Since then, they supported patenting activities with various incentives, and by setting new, more patenting friendly, regulations regarding the examination of patent applications. Moreover, China’s high-tech companies and telecoms have become significant global players, not only conducting patenting activities but also buying patent rights. State Grid Corporation of China, which is in top 100 patent applicants worldwide, leads in patenting activities regard big data and prediction. Specifically, State Grid Corporation took ninth place when it comes to the application of patent families for the period between 2011 and 2014, especially for the following technological fields: electrical machinery, apparatus and energy, as well as technical content related to measurement.

Stakeholders who are interested in harnessing big data analytics solutions can choose between numerous vendors. However, vendors often acquire patents’ rights, so they do not have to be patent assignees or inventors. Instead, they make strategic investments in patents, acquire patents and manage patent portfolios, which allows them to focus on their core activities and provide innovative solutions to clients. For example, in 2015, Avigilon Corporation, a global provider of surveillance solutions, including video analytics, acquired 126 USA and international patents from VideoMining Corporation, FaceDouble Incorporated, Behavioral Recognition Systems and ITS7 Pty. The total value of patents was US$135,375,000, covering technical content: different video analytics capabilities such as behavioral analysis, in-store object tracking, video segmentation, anomaly detection, image classification, as well as patents related to programming of remote security camera and network camera system.

*What?*

Search revealed the most frequent patent topics are related to technological solutions (G06Q - Analogue computers), data processing (G06F - Electrical digital data processing), and specific areas (G06N - Computer systems based on specific computational models). This finding is in line with the specific challenges related to big data identified by Sivarajah et al. (2017): (i) data challenges that are related to the characteristics of big data, (ii) process challenges, including challenges related to big data analysis and modelling, and (iii) management challenges that cover privacy, security, data governance, data and information sharing, cost and operational expenditure and data ownership challenges.

Number of patent families focus to technological solutions and data processing solutions, which try to solve specific challenges related to big data analytics. Techniques of predictive analytics can be divided into two group (Gandomi and Haider, 2015): (i) techniques for discovering historical patterns and extrapolating an outcome variable(s), and (ii) techniques for exploring the interdependencies between outcome and explanatory variables. Predictive analytics mostly relies on statistical techniques. However, while the conventional statistical methods rely on statistical significance to examine a significance of the specific relationship, big data analysis is often conducted on majority or entire population, so statistical significance is not that important for big data as compared to small samples of a population. Furthermore, when it comes to computational challenges, many conventional methods for small samples do not scale up to big data. For that reason, existing methods are extended and modified for parallel and distributed tasks. Additionally, big data unique characteristics cause some problems when it comes to estimating predictive models for big data (Gandomi and Haider, 2015): noise accumulation, spurious correlation and incidental endogeneity. Noise accumulation or accumulated estimation error sometimes results in overlooking some significant variables. Spurious correlation appears when some variables appear to be correlated because of high dimensionality of big data. In addition, incidental endogeneity, the dependence of the error term and variables, is common in big data. Extreme machine learning techniques have been extended for tasks such as clustering and adapted for parallel computation, which makes them feasible for big data analytics (Huang et al., 2014). Zhang et al. (2018) discussed the role and future of deep learning techniques in big data analytics that are used for image, audio and text analytics. Another issue of big data analytics is related to big data proneness to noise, outliers, inconsistencies and incompleteness (Wu, X., Zhu, Wu, G.-Q. and Ding, 2014). Additionally, re-utilization of existing big data should be taken into account. Most of the big data analytics algorithms will be designed to support parallel and distributed computing. This raise problem regard bottlenecks of algorithms that may occur because of synchronization and information exchange issues (Tsai et al., 2015; Wu et al., 2014). Additionally, big data technology needs improvements regard efficiency of format conversion of heterogeneous data, big data transfer and performance of real-time processing of big data.

Identified association rules indicate some specific domains of their usage such as market research, buyer profiling, price estimation or determination, computer-aided design and so on. Text mining revealed that following topics occur together: (i) real-time systems focusing to e.g. weather forecasting, (ii) database technologies related to preprocessing of specific data, such as big traffic data, (iii) technical challenges, such as cloud computing, (iv) specific topics, such as monitoring public opinion, and analyzing behavior data, (v) methodological challenges, such as usage of support vector machine to increase prediction accuracy, and (vi) specific topics related to information extraction from historical data, e.g. risk assessment.

Some of these topics indicate patenting activities for challenges related to data management and data integration. Safety and privacy are always key challenges and concerns when it comes to information and communication technology, as well as data. Security-related big data challenges are big data privacy, safety and big data application in information security (Chen et al., 2014). Big data privacy includes protection of personal privacy during data handling. Nowadays, usage of information and communication technology potentiate easy and simple generation and acquisition of large amounts of users’ data. Hence, it is highly important for users to raise their awareness on which of their personal data third parties collect and how it is used. Big data safety mechanisms, such as efficient cryptography of big data and schemes for safety management, are under development. Efficiency of big data mechanisms is assured only if data availability, completeness, controllability, traceability and confidentiality are enabled (Chen et al., 2014).

**CONCLUSION**

Big data will influence society, economy and it will drive the progress of technologies in the near future. It causes fusion of different disciplines, which is particularly visible when it comes to big data analytics. Big data influence operations and decision making in various application fields. On the other hand, society promotes the progress of technologies, including widespread usage and development of big data. Additionally, big data encouraged fusion of different technologies, such as the Internet of Things, cloud computing and so on, and forces exploration of new and innovative technologies for handling big data. People are participants of big data, both users and generators of big data. Generation of real-time and streaming data, online network data, Internet of Things and mobile data, geography data (e.g. geo-tag or location-based real-time geographic data), spacial-temporal data, and visual data represent trends in big data area (Lv et al., 2017; Brown et al., 2011). Shortly, it is expected that the volume of such data will grow to a large degree due to technological advances and development in related areas, such as geo-databases or wireless sensor networks. Furthermore, demands from a wide range of application areas, along with new database and processing technologies, drive the modification of existing techniques and development of new techniques for big data analytics.

The chapter presents a patent analysis technical area of big data for prediction based on data searched and gathered from PatSeer patent database. Authors analyzed 296 active simple patent families related to big data for prediction assigned from 2013 to October 2017. The patent analysis was conducted in four stages related to (i) the patent search and selection, (ii) timeline, geographic origin and patents assignees analysis, (iii) patents according to IPC system patent area, and (iv) text mining patent analysis. An analysis of the 296 simple patent families related to big data for prediction was conducted to achieve the goal of this research.

The analysis provided insights into the technical area of big data for prediction. The increasing trend is in patenting the technical content of big data for prediction is present from 2013, with 122 simple patent families registered in 2016 and 107 simple patent families registered in 2017, until October. This is due to an increasing interest in big data and new opportunities big data brings. Authors revealed that the patenting activities related to big data for prediction are spread across China and South Korea which organizations assigned the majority of patents related to the technology of interest. The organization that registered the largest number of simple patent families related to big data for prediction in the observed period is State Grid Corporation registered in China (227 simple patent families or 7.43%). Other organizations that registered a larger number of simple patent families are companies such as Business Big Data, NA Computer Network Information Security MAN, Shanghai Fuli Information Technology and academic institutions such as Hohai University, Beijing Jiaotong University, University of South China and so on.

Next, the protected technical content of big data for prediction simple patent families was analyzed by using the International Patent Classification system at the section, class, sub-class, main group or sub-group level. The simple patent families were mostly registered under the section G codes (474 times) with the following classes most frequently assigned: G06 - Computing; calculating; counting instruments (407 times), G08 – Signaling instruments (29 times) and G01 – Measuring; testing instruments (24 times). Therefore, computing instruments have been the major focus of assignees-inventors. Specifically, a significant number of simple patent families were information retrieval, database structures or file system as a part of data processing systems specially adapted for administrative, commercial, financial managerial, supervisory or forecasting purposes.

Furthermore, association rules analysis revealed rules that are trivial due to dataset limitations. For better results, weighted association rules should be applied in future research with additional patent data such as backward citations and the number of IPC codes. Therefore, authors conclude that the technical content of the observed simple patent families is not heterogeneous, but association rules indicate some specific domains.

Finally, authors used the phrase extraction process combined with the cluster analysis to detect most common topics appearing in big data for prediction simple patent families’ abstracts. Most frequent phrases occurring in big data for prediction simple patent families’ abstracts were real time, data mining, early warning and neural networks. The phrases that occur the most often with the phrase real-time are mostly related to data analysis such as historical data, management systems, real-time performance and monitoring data (Belfo et al., 2015). The phrases that occur the most often with the phrase data mining are mostly related to the phrase historical data, and methods and techniques of data analysis such as machine learning, natural language or deep learning. The phrases that occur the most often with the phrase early-warning indicate specific purposes of the weather forecast and power supply domain, and source of data analyzed by early-warning systems such as monitoring data, weather information, environment information. The phrase neural network occurs with phrases that indicate its specific applications areas such as power load, behavior analysis, weather forecast, feature extraction or medical diagnosis. Cluster analysis identified 6 groups of topics regard big data for prediction patents and the connections between keywords – phrases are visualized by using a network graph to explore relationships, to detect underlying patterns and structures of co-occurrences.

**REFERENCES**

Akter, S., and Wamba, S.F. (2016). Big data analytics in E-commerce: a systematic review and agenda for future research. Electron Markets, 26, 2016, 173–194.

Altunas, S., Dereli, T., and, Kusiak, A. (2015). Analysis of patent documents with weighted association rules, Technological Forecasting and Social Change, 92, 249-262.

Belfo, F., Trigo, A., & Estébanez, R. P. (2015). Impact of ICT Innovative Momentum on Real-Time Accounting. Business Systems Research Journal, 6(2), 1-17.

Boyd, D., and Crawford, K. (2012). Critical questions for big data, Communication and Society, 15(5), 662-679.

Bradlow, E.T., Gangwar, M., Kopalle, P., and Voleti, S. (2017). The Role of Big Data and Predictive Analytics in Retailing. Journal of Retailing, 93(1), 79-95.

Brown, R. A., & Sankaranarayanan, S. (2011). Intelligent store agent for mobile shopping. International Journal of E-Services and Mobile Applications (IJESMA), 3(1), 57-72.

Brügmann, S., Bouayad-Agha, N., Burga, A., Carrascosa, S., Ciaramella, A., Ciaramella, M., Codina-Filba, J., Escorsa, E., Judea, A., Mille, S., Müller, A., Saggion, H., Ziering, P., Schütze, H., and Wanner, L. (2015). Towards content-oriented patent document processing: Intelligent patent analysis and summarization, World Patent Information, 40, 30-42.

Chen, H., Chiang, R. H. L., and Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly, 36(4),1165-1188.

Chen, M., Mao, S., and Liu, Y. (2014), Big Data: A Survey. Mobile Networks and Applications, 19(2), 171-209.

Dmitriyev, V., Mahmoud, T., & Marín-Ortega, P. M. (2015). SOA enabled ELTA: approach in designing business intelligence solutions in Era of Big Data. International Journal of Information Systems and Project Management, 3(3), 49-63.

Everitt, B.S., Landau, S., Leese, M., and Stahl, D. (2011). Hierarchical clustering. In: Shewhart, W.A., Wilks, S.S. (eds). Cluster Analysis, 5th Edition, Wiley Series in Probability and Statistics. Ed., John WileyandSons, Ltd.

Gandomi, A. and Haider, M. (2014). Beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management, 35, 2015, 137 – 144.

Garret, M.A. (2014). Big Data analytics and cognitive computing – future opportunities for astronomical research. IOP Conference Series: Materials Science and Engineering, 67, 012017.

Gepp, A., Linnenluecke, M.K., O’Neill, T.J. and Smith,T. (2018). Big data techniques in auditing research and practice: Current trends and future opportunities. Journal of accounting literature, 40, 2018, 102-115.

Günther, A.W., Rezazade, M. H., Huysman, M., and Feldberg, F. (2017). Debating big data: A literature review on realizing value from big data. The Journal of Strategic Information Systems, 26(3), 191-209.

Han, Q., Heimerl, F., Codina-Filba, J., Lohmann, S., Wanner, L., and Ertl, T. (2017). Visual patent trend analysis for informed decision making in technology management. World Patent Information, 49, 34-42.

Huang, G., Huang, G.-B., Song, S. and You, K. (2014). Trends in extreme machine learning: A review. Neural Networks, 61, 2015, 32-48.

Ji, W., and Wang, L. (2017). Big data analytics based fault prediction for shop floor scheduling. Journal of Manufacturing Systems, 43(1), 187-194.

Kim, G., and Bae, J. (2017). A novel approach to forecast promising technology through patent analysis. Technological Forecasting and Social Change, 117, 228-237.

Kim, M., Park, Y., and Yoon, J. (2016). Generating patent development maps for technology monitoring using semantic patent-topic analysis, Computers and Industrial Engineering, 98, 289-299.

Kyebambe, M., Cheng, G., Huang, Y., He, C., and Zhang, Z. (2017). Forecasting emerging technologies: A supervised learning approach through patent analysis. Technological Forecasting and Social Change, 125, 236-244.

Lee, J., Ardakani, H.D., Yang, s., and Bagheri, B. (2015). Industrial Big Data Analytics and Cyber-ph

Lv, Z., Song, H., Basanta-Val, P., Steed, A. and Jo, M. (2017). Next-Generation Big Data Analytics: State of the Art, Challenges, and Future Research Topics. IEEE Transactions on Industrial Informatics, 13(4), 1891-1899.

Madani, F., and Weber (2016). The evolution of patent mining: Applying bibliometrics analysis and keyword network analysis,World Patent Information, 46, 32-48.

McAfee, A., and Brynjolfsson, E. (2012). Big data: the management revolution. Harvard Bussiness Review, 90(10), 60-68.

Miah, S.J., Vu, Q.H., Gammack, J., and McGrath, M. (2017). A Big Data Analytics Method for Tourist Behavior Analysis. Information and Management, 54(6), 771-785.

Niemann, H., Moehrle, M. G., and Frischkorn, J. (2017). Use of a new patent text mining and visualization method for identifying patenting patterns over time: Concept, method and test application, Technological Forecasting and Social Change, 115, 210-220.

NIST Big Data Public Working Group (2017). Big Data Interoperability Framework: Volume 1, Definitions [online]. Accessed at: http://bigdatawg.nist.gov/home.php [5th November 2017]

Parr Rud, O. (2011). Invited article: Adaptability. Business systems research journal: international journal of the Society for Advancing Business & Information Technology (BIT), 2(2), 4-12.

PatSeer. (2017). <http://patseer.com/> [Accessed 13/10/2017].

Pejić Bach, M., Pivar, J., & Dumičić, K. (2017). Data anonymization patent landscape. Croatian Operational Research Review, 8(1), 265-281.

Rassam, M.A., Maarof, M.A. and Zainal, A. (2017). Big Data Analytics Adoption for Cyber-Security: A Review of Current Solutions, Requirements, Challenges and Trends. Journal of Information Assurance and Security, 12(4). 124-145.

Salehan, M., and Kim, D.J. (2016). Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics. Decision Support Systems, 81(January 2016), 30-40.

Sharma, P. and Tripathi, R.C. (2017). Patent citation: A technique for measuring the knowledge flow of information and innovation, World Patent Information, 51, 2017, 31-42.

Sharma, R., and Kankanhalli, A. (2014). Transforming decision-making processes: a research agenda for understanding the impact of business analytics on organizations. European Journal of Information Systems, 23 (4), 433-441.

Shirdastian, H., Laroche, M., and Richard, M.O. (2017). Using big data analytics to study brand authenticity sentiments: The case of Starbucks on Twitter. International Journal of Information Management, in Press, Corrected Proof, ISSN 0268-4012, <https://doi.org/10.1016/j.ijinfomgt.2017.09.007>.

Sinha, M., and Pandurangi, A. (2015). Guide to Practical Patent Searching And How To Use Patseer For Patent Search And Analysis. Gridlogics Technologies, Pune.

Sivarajah, U., Kamal, M.M., Irani, Z., and Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. Journal of Business Research, 70, 2017, 263-286.

Song, K., Kim, K., and Lee, S. (2018). Identifying promising technologies using patents: A retrospective feature analysis and a prospective needs analysis on outlier patents. Technological Forecasting and Social Change, 128, 118-132.

Tsai, C.W. , Lai, C.F., Chao, H.C, and Vasilakos, A.V. (2015). Big data analytics: a survey. Journal of Big Data, 2:21

Vera-Baquero, A., Colomo-Palacios, R., Molloy, O., & Elbattah, M. (2015). Business process improvement by means of Big Data based Decision Support Systems: a case study on Call Centers. International Journal of Information Systems and Project Management, 3(1), 5-26.

Waller, M.A., and Fawcett, S. E. (2013). Data Science, Predictive Analytics, and Big Data: A Revolution that Will Transform Supply Chain Design and Management. Journal of Business Logistics, 34(2), 77-84.

World Intellectual Property Organization - WIPO (2017). Guide to the International Patent Classification. Accessed at: <http://www.wipo.int/export/sites/www/classifications/ipc/en/guide/guide_ipc.pdf> [Accessed 6/11/2017]

World Intellectual Property Organization, Economics and Statistics Division (2016). World Intellectual Property Indicators 2016. Accessed at: <http://www.wipo.int/edocs/pubdocs/en/wipo_pub_941_2016.pdf> [Accessed 30/05/2018]

Wu, X., Zhu, X., Wu, G.-Q., and Ding, W. (2014). Data Mining with Big Data. IEEE Transactions on Knowledge and Data Engineering, 26(1), 97-107.

Zhang, Q., Yang, L.T., Chen, Z. and Li, P. (2018). A survey on deep learning for big data. Information fusion, 42, 146-157.